A Baseline for Content-Based Blog Classification

Olof Görnerup
Swedish Institute of Computer Science (SICS)
Box 1263
SE-16429 Kista, Sweden
olofg@sics.se

Magnus Boman
Swedish Institute of Computer Science (SICS)
Box 1263
SE-16429 Kista, Sweden
mab@sics.se

ABSTRACT

A content-based network representation of web logs (blogs) using a basic word-overlap similarity measure is presented. Due to a strong signal in blog data the approach is sufficient for accurately classifying blogs. Using Swedish blog data we demonstrate that blogs that treat similar subjects are organized in clusters that, in turn, are hierarchically organized in higher-order clusters. The simplicity of the representation renders it both computationally tractable and transparent. We therefore argue that the approach is suitable as a baseline when developing and analyzing more advanced content-based representations of the blogosphere.

1. INTRODUCTION

For any given blog, a number of other blogs is related to it through content overlap. The owner of the blog and its readers—or indeed anyone interested in blog navigation in general—are probably interested in learning more about those related blogs. The problem is that the sheer size of the blogosphere makes keeping up with the dynamic set of related blogs next to impossible. To assist, a vast array of tools and algorithms have been presented (cf. [1; 2]). While it is important to remember that blog networks differ from general Web page networks in several important respects [1], much of the mathematics governing the latter is reusable for modeling also the former (see, e.g., [9]). The classification of blogs can be made according to different criteria, including blog entry similarity [14], interblog communication and community stability [8], sense of community among bloggers [4], discussion keyword correlation [3], and a host of machine learning and statistics approaches. To date, however, almost all these tools and algorithms require human intervention and considerable time investment to overcome problems with bootstrapping, tuning, and not least semantics. Understanding a graph, perhaps with thousands of vertices and edges, pertaining to describe relevance to one's own blog according to some set of possibly esoteric or advanced criteria is not straightforward. We address this problem by presenting a method for generating a network of relevant blogs by means of the simplest similarity criterion there is: word overlap. We will demonstrate that even this naïve approach allows us to accurately classify blogs with respect to their contents, filter out spam weblogs (splogs) and multiple occurrences (i.e., blogs linked to from multiple URLs) in addition to providing a global overview of the blogosphere. The procedure is transparent, modular, computationally efficient, and requires no human monitoring or control. Using a network representation also enables one to employ a wealth of theory and techniques recently developed in network theory [12].

We describe our methodology in the section that follows. Section 3 presents results of our experiments, using data from the Swedish blogosphere. We conclude the paper by discussing the results and pointing to possible future directions in Section 4.

2. METHODOLOGY

Our overall approach is to provide a global view of the blogosphere in the form of a network, where vertices constitute blogs and where weighted edges constitute similarity relations between blogs. There is a link between two blogs if they are related, and the strength of the link is given by the degree of similarity.

2.1 Similarity measure

To estimate the similarity between blogs we simply compare the overlap of occurring words. In more formal terms, the similarity is quantified in the following way. Given two blogs i and j, let $\mathcal{W}_i$ denote a set of words (to be qualified below) that occur in i and $\mathcal{W}_j$ a set of words that are used in j. The similarity s_{ij} between i and j is then defined as

$$s_{ij} = \frac{|\mathcal{W}_i \cap \mathcal{W}_j|}{|\mathcal{W}_i \cup \mathcal{W}_j|}. \quad (1)$$

In other words, s_{ij} is the fraction of all words in $\mathcal{W}_i$ and $\mathcal{W}_j$ that are shared by the two sets. It holds that $0 \leq s_{ij} \leq 1$, where $s_{ij} = 1$ if $\mathcal{W}_i$ and $\mathcal{W}_j$ are identical and $s_{ij} = 0$ if they do not share a single word. The similarity measure is equivalent to Tversky's Ratio model [15], which has been found to be a good trade-off between simplicity and performance among text document similarity models [10].

We do not consider the full word sets of blogs—literally *all* occurring words—for several reasons. Firstly, comparing very common words ("the", "it", "do", etc.) will only provide a negligible amount of similarity information. The use of very uncommon words, on the other hand, is likely to tell us a lot about the characteristics of a blog. However, at the same time we do not want to consider words that are *too* uncommon—for instance occurring only a handful of times in the blogosphere during the course of several months—since these are often misspellings and typos that only add noise to the statistics. Another reason for not considering all words is a pragmatic one. Analyzing tens of thousands of

blogs can be computationally expensive. By utilizing Zipf's law [16], which implies that a few of the most common words stand for a large majority of word occurrences[1], the computational cost is drastically reduced.

2.2 Network structure

The global structure of a similarity network provides valuable information that facilitates an exploration of the blogosphere. Specifically, blogs are organized in clusters that reflect domains of topics such as politics, books, technology, or music. What characterizes clusters is that there are significantly higher densities of edges within clusters than between them. Such clusters are referred to as *communities* in the network literature. We chose to use the term cluster here, however, due to the ambiguity of the word "community" in this context. A cluster may indeed constitute a community in the social meaning of the word, although it is not necessarily so. In fact, one advantage of using a content-based similarity measure is that it relates blogs that otherwise lack explicit (social or hyper-) links.

Clusters can be quantified as follows [11]. Let $\{v_1, v_2, ..., v_n\}$ be a partition of a set of vertices into n groups, r_i the degree of edge weights (i.e., similarities) internal to v_i (the sum of internal weights over the sum of all weights in the network) and s_i the degree of weights of edges that start in v_i. The degree of cluster structure is then defined as

$$Q = \sum_{i=1}^{n} (r_i - s_i^2). \quad (2)$$

To infer clusters in the blog network we employ Clauset's method [5], which aims to find cluster assignments—a partition of the set of vertices—such that Q is maximized.

Edges are not only organized in clusters, but the clusters are in turn organized in higher order clusters. This hierarchical organization of the graph is interesting since it may enable automatic generation of hierarchical blog taxonomies. Here we use a Monte Carlo sampling technique by Clauset et al. [6] to identify hierarchical structure in the blog network.

2.3 Case study: The Swedish blogosphere

We have tested our approach on the Swedish blogosphere. The API of the blog search engine *Twingly*[2] was used to collect data, where individual blog posts from the period March-July 2009 were fetched and indexed. In the spirit of keeping things simple, we refrained from applying ad hoc textbook pre-processing, such as stemming words or removing slang or acronyms (with respect to predefined word lists) [2]. Instead we relied on basic word frequency statistics to filter out words in the following specific way. During indexing, word frequencies were calculated. We then discarded all words occurring less than ten times. Of the remaining words we kept those that occured in the fifth percentile of the frequency distribution. For each blog, we collected its set of those occurring words. Blogs that had word sets of size 25 or larger were kept. This ensured a meaningful similarity measure and also filtered out a considerable amount of spam blogs. At this point, 21564 blogs remained. The similarities between each pair of these were calculated, and stored if equal to or above 0.025. We have tested to vary the above parameters and the results reported here appear to be stable.

[1]More specifically, the frequency of a word is inversely proportional to its rank; $f_n \sim 1/n^a$, where n is the rank ($n = 1$ for the most common word, $n = 2$ for the second most common word, etc.) and a is some exponent.

[2]http://www.twingly.com/

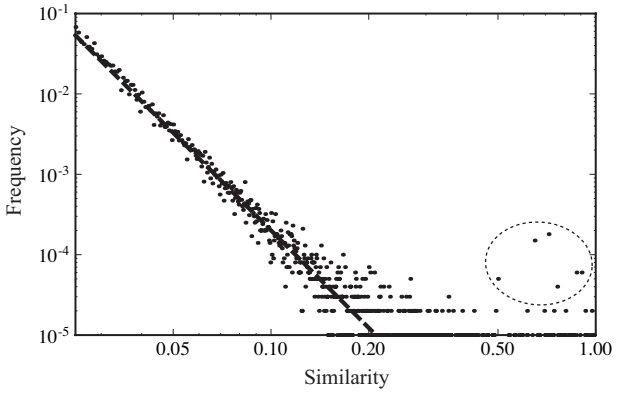

Figure 1: Log-log plot of the similarity distribution of blogs. The dashed line denotes a least-squares linear regression in the region $[0.025, 0.2]$ which has a slope of about -4.0. Despite the simplicity of the similarity measure, it provides valuable information: Similarities between blogs that occur in multiple copies (or nearly copies) appear in the very end of the tail, and splogs are easily identified by outliers (enclosed by a dashed ellipse).

3. RESULTS

The distribution of similarities is shown in Fig. 1. Interestingly, remaining splogs are revealed by outliers; similarities that are "unusually" high. Groups of multiple occurrences of the same (or slightly deviating) blogs are also easily identified in the tail of the distribution. The acquired network (Fig. 2) displays a distinct clustered structure. Splogs are also revealed here by a tightly connected cluster. Fig. 3 shows automatically inferred clusters when similarities $\gamma = 0.05$ are considered. When manually sampling and classifying blogs, clusters and blog classes are found to be consistent. An example of the hierarchical organization of the blog network is depicted in Fig. 4 in the form of a dendrogram of a "food and beverages" cluster.

4. DISCUSSION AND OUTLOOK

We have shown that the signal in raw blog data is so strong that even our basic similarity measure—word occurrence overlap—is capable of capturing valuable structural information. Since the measure is computationally tractable it therefore enables an efficient mean for classifying blogs when used in concurrence with fast graph clustering algorithms. We grant that there are more advanced—and possibly more accurate—(document) similarity measures, e.g., Latent semantic analysis [7] and Random indexing [13]. However, we believe that the minimal (non-trivial) measure employed here is suitable as a baseline when studying blog similarity networks. The measure is admittedly simplistic, yet this is also its strength since it decreases the risk of causing hidden representation-dependent artifacts that are more difficult to identify when using more advanced similarity measures.

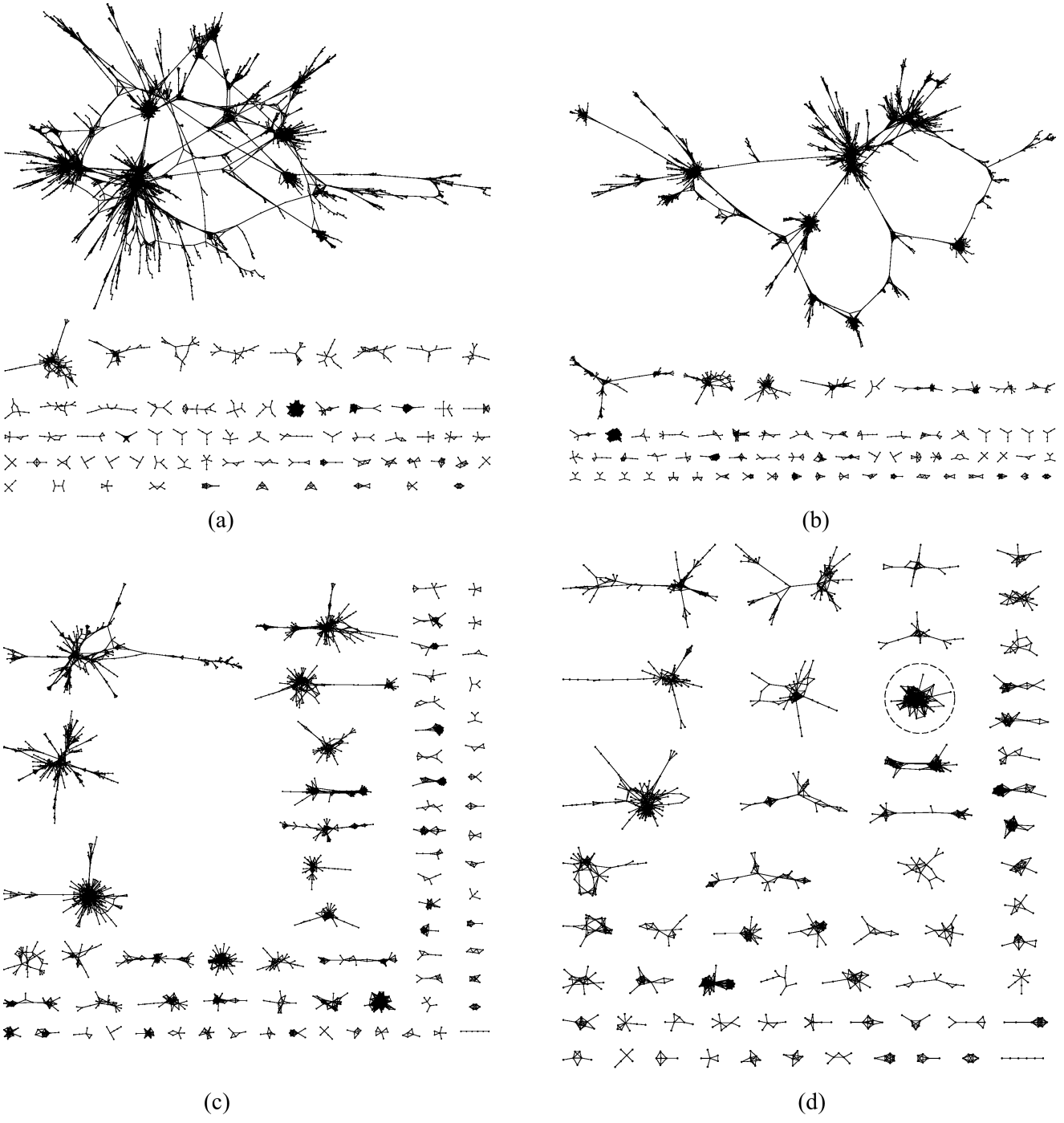

Figure 2: Visualization of the Swedish blogosphere, where blogs with similarities $\geq \gamma$ are shown. (a) $\gamma = 0.04$. (b) $\gamma = 0.045$. (c) $\gamma = 0.055$. (d) $\gamma = 0.07$. A spam blog cluster is enclosed within a dashed circle.

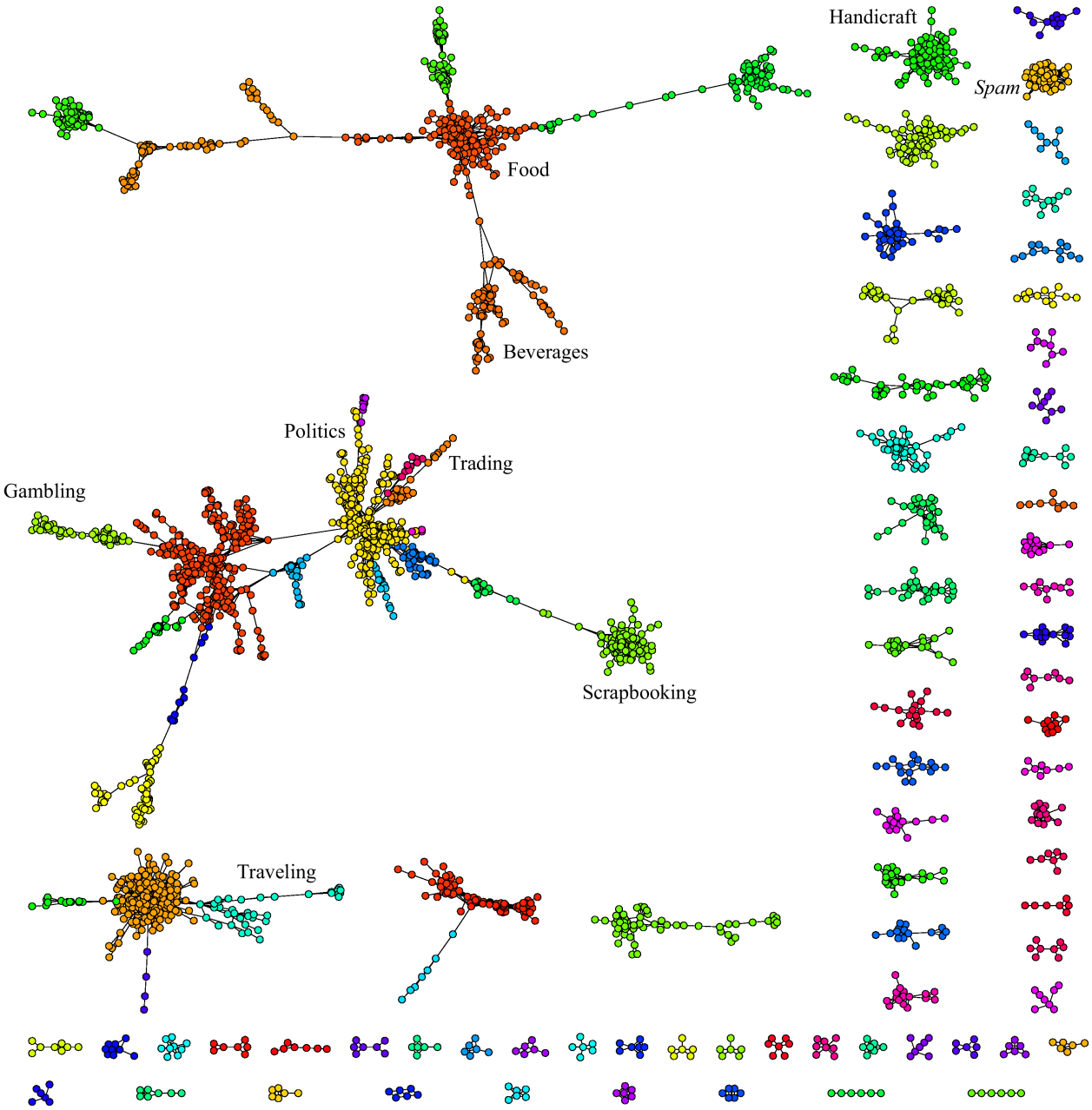

Figure 3: Inferred categories of Swedish blogs ($\gamma = 0.05$). Some examples are marked with text.

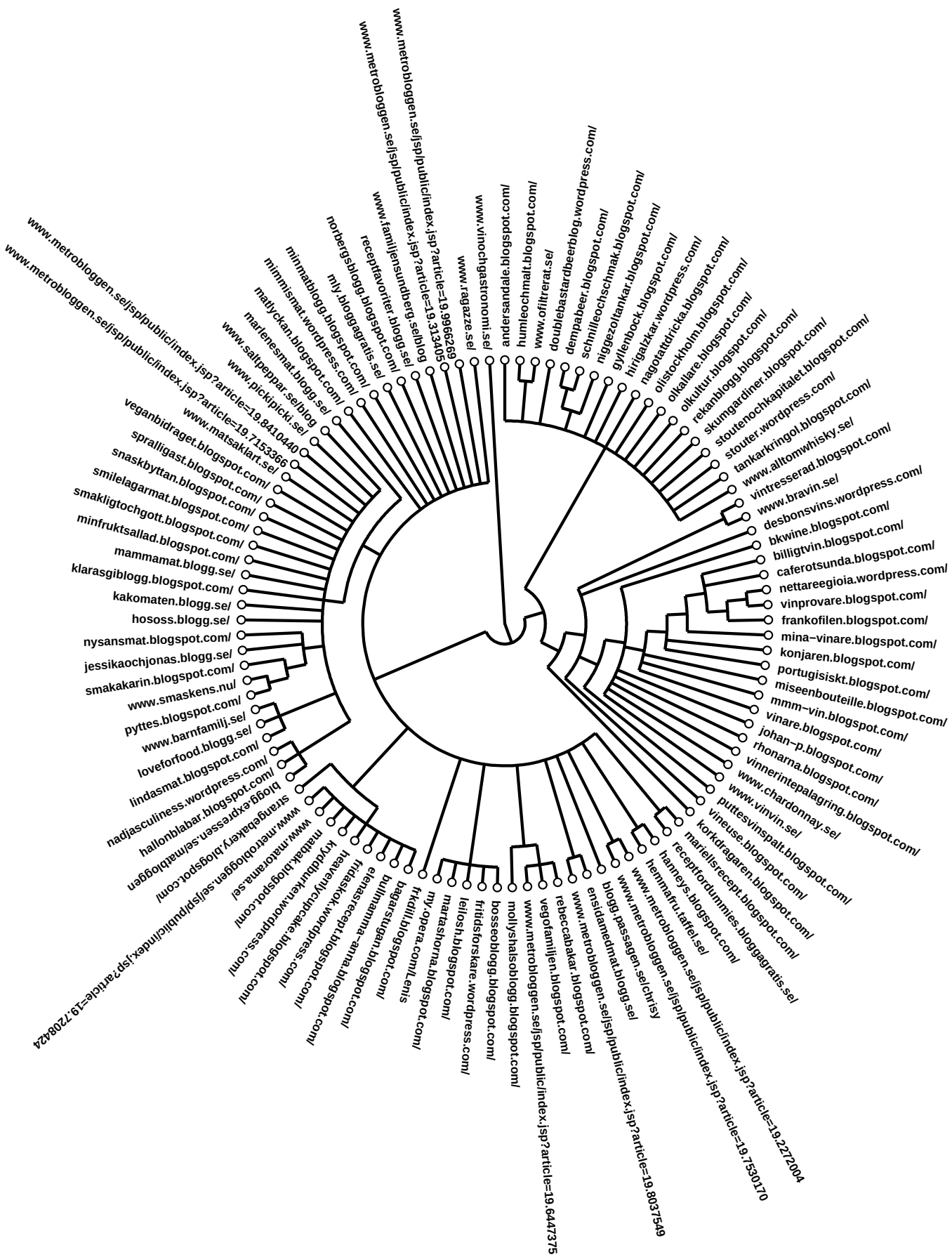

Figure 4: Dendrogram of the food-and-beverages cluster. Blogs are in turn organized in separate vine, beer and food clusters. Leaf nodes are labeled with corresponding blog URLs.

An issue that needs to be addressed in future work is that of validation. How can we know that acquired blog classifications are correct? So far, our approach has been to examine a random sample of blogs and subjectively confirm that their contents is consistent within inferred blog clusters. Such empirical evaluations can be problematic, however. In some cases, a manual classification is clear cut (i.e. identifying that two blogs that solely treat Belgian beer belong to the same class), but not always. A more quantitative measure that verifies the result is therefore desirable. This can also be turned into an epistemological question. One can for example imagine cases when the blog classes acquired from the similarity network can be used to evaluate *other* blog classifications (including our own subjective one). However, in this discussion we have more pragmatic and application-oriented evaluation methods in mind.

We have treated only a few structural aspects of the blog network here. These deserve more attention, as well as network dynamics and evolution: How does information diffuse and change in the network, and how does the network structure itself change over time? For instance, through an analysis along these lines one may perhaps trace how emerging trends or news proliferate in and between specific topic domains of the blog similarity network.

Another possible future direction that we have only touched on briefly here concerns splog detection. We have seen that blog classification also groups together splogs. If an individual blog is identified as a splog (e.g., by examining the distribution of blog similarities), it is likely that its associated blog cluster also consists of splogs. If such a relation proves to hold true in general, it enables splog detection and removal at the level of blog clusters rather than individual blogs, which presumably would be much more efficient.

Acknowledgements

OG was funded by The Internet Infrastructure Foundation (.SE). The authors thank Twingly for providing blog data and Aaron Clauset for sharing source code for the hierarchical structure inference algorithm and for the radial dendrogram visualization script used for rendering Fig. 4.

5. REFERENCES

[1] N. Agarwal and H. Liu. Blogosphere—research issues, tools, and applications. *SIGKDD Explorations*, 10(1):18–31, 2008.

[2] N. Agarwal and H. Liu. *Modeling and Data Mining in Blogosphere*. Morgan and Claypool Publishers, 2009.

[3] N. Bansal, N. Koudas, F. Chiang, and F. W. Tompa. Seeking stable clusters in the blogosphere. In *Proceedings of the 33rd international conference on Very large data bases*, pages 806–817. VLDB Endowment, 2007.

[4] A. Chin and M. Chignell. A social hypertext model for finding community in blogs. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 11–22. ACM, New York, 2006.

[5] A. Clauset. Finding local community structure in networks. *Physical Review E*, 72:026132, 2005.

[6] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[8] Y. C. et al. Structural and temporal analysis of the blogosphere through community factorization. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 163–172. ACM, New York, 2007.

[9] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[10] M. D. Lee, B. Pincombe, and M. Welsh. An empirical evaluation of models of text document similarity. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1254–1259. Erlbaum, 2005.

[11] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 2004.

[12] M. E. J. Newman, A. L. Barabási, and D. J. Watts, editors. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[13] J. K. Pentti Kanerva and A. Holst. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 103–6, 2000.

[14] C. Tauro, S. Ahuja, M. A. Prez-Quiones, A. Kavanaugh, and P. Isenhour. Vizblog: Discovering conversations in the blogosphere. In *Technology demonstration at Directions and Implications of Advanced Computing - Conference on Online Deliberation*, University of California, Berkeley, 2008.

[15] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

[16] G. K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison Wesley, Cambridge MA, 1949.